Identification des motifs séquentiels affectant l'attrition des employés

Youssef Oubelmouh[*,**], Frédéric Fargon[*], Cyril de Runz[**], Arnaud Soulet[**], Cyril Veillon[*]

[*]Devoteam, Levallois-Perret, prenom.nom@devoteam.com
[**] BDTLN, LIFAT, Université de Tours, Blois, prenom.nom@univ-tours.fr

Résumé. Ce travail est une version courte d'un article publié à DSAA (Oubelmouh et al., 2023a) concernant la problématique de l'attrition des employés au sein des organisations. Alors que réduire l'attrition est devenu un enjeu majeur dans les entreprises, la littérature sur ce sujet est relativement limitée en comparaison avec celle concernant l'attrition des clients. De plus, même les études abordant ce thème ne prennent souvent pas en compte l'impact du temps et de la durée sur les taux d'attrition. Dans ce contexte, cette recherche combine deux approches pour combler cette lacune : l'exploration de motifs fréquents dans les séquences d'événements et l'analyse de survie avec Kaplan-Meier. En introduisant la notion de *motifs séquentiels impactant la survie*, cette étude identifie les événements ayant un impact significatif sur l'estimateur de survie. Les résultats suggèrent que certains motifs ainsi que l'ajout d'événements spécifiques peuvent influencer positivement ou négativement la rétention des employés.

1 Introduction

L'attrition des employés est un problème croissant dans les entreprises technologiques du monde entier, et en particulier dans les entreprises informatiques/cabinets de conseil. Cependant, les travaux récents dans le domaine de l'analyse de données ne prennent pas en compte la notion du temps et des durées dans l'étude de l'attrition des employés (Oubelmouh et al., 2023b). Pour mieux saisir l'aspect temporel de l'attrition, nous voudrions combiner les techniques d'exploration de motifs séquentiels dans les séquences d'événements et l'analyse de survie. D'une part, l'exploration de motifs séquentiels vise à découvrir des motifs significatifs dans des séquences d'événements, telle que la séquence des changements subis par un employé. Ce domaine actif fait l'objet de nombreuses études (Mooney et Roddick, 2013; Truong-Chi et Fournier-Viger, 2019). D'autre part, l'analyse de survie est utilisée pour modéliser les données relatives à la durée des événements, comme le temps écoulé entre l'embauche et la démission, et pour estimer la fonction de survie d'une population. Wang et al. (2019) ont mené une enquête exhaustive sur les diverses méthodes statistiques et non statistiques utilisées dans l'analyse de survie. Il existe quelques travaux combinant l'exploration de motifs et

l'analyse de survie (Ritschard et al., 2008; Ren et al., 2019; Mattos, 2021), mais ils ne s'appuient pas sur des motifs modifiant la survie. Par ailleurs, aucun travail n'exploite ce concept pour l'analyse de l'attrition des employés.

Dans cet article, notre objectif est de proposer une approche permettant d'extraire les facteurs d'attrition/de rétention des employés en tenant compte de l'aspect temporel du phénomène grâce à l'exploration de motifs séquentiels et à l'analyse de survie. Notre approche vise à répondre aux questions suivantes :

1. Quel est l'enchaînement des événements qui conduisent à une démission ?
2. Quels sont les facteurs permettant de retarder ou accélérer la démission d'un employé à partir d'un contexte donné ?
3. Comment relier l'exploration de motifs séquentiels à l'analyse de survie ?

Pour répondre à ces questions, nous définissons la notion de courbe de survie de Kaplan-Meier pour un motif séquentiel. Nous introduisons ensuite les motifs *impactant la survie* qui modifient significativement l'aire sous la courbe de survie par rapport à un contexte donné. Nos expérimentations identifient des motifs pertinents pour remédier à l'attrition.

2 Préliminaires

Définitions pour les motifs séquentiels Soit $\mathcal{A} = \{A_1, A_2, ..., A_n\}$ un ensemble d'attributs. Soit $\mathcal{C} = \{C_1, C_2, ..., C_n\}$ représentant l'ensemble des changements. Un changement correspond à la variation d'un attribut. Par exemple, l'attribut A_1 passant d'un état à un autre est un changement d'attribut (C_1 a lieu). On définit $\mathcal{E}$ un ensemble d'événements, où un événement est une paire contenant un changement et un temps. Donc, $\mathcal{E} \subseteq \mathcal{C} \times \mathbb{N}$. Par exemple, $E_4 = (C_4, t_1)$ est un événement, indiquant que l'attribut A_4 a changé d'état au temps t_1. Une séquence d'événements $Seq \in \mathcal{S}$ est un ensemble d'événements ordonnés dans le temps, tel que $Seq_1 = \langle E_3, E_8, ..., E_2 \rangle = \langle (C_3, t_0), (C_8, t_1), ..., (C_2, t_f) \rangle$, où $\forall i$, $E_i \in \mathcal{E}$, et pour tout t_i, t_j, $i < j \Rightarrow t_i \leq t_j$. Lorsque le contexte est clair, Seq_1 peut désigner $< C1, \ldots, Cn >$.

Pour illustrer nos définitions, le tableau 1 fournit un jeu de données contenant 8 séquences décrites par les changements $\{S, M, P, A, B, C, D, E, F, G, H, J\}$ où nous avons supprimé les temps par souci de simplicité en ne conservant que la durée entre l'embauche et la démission. Nous avons toutefois besoin des durées entre l'embauche et la démission pour pouvoir appliquer notre méthodologie.

Un motif (séquentiel) X est une séquence ordonnée de changements (par exemple, $\langle S, M, P \rangle$). $\mathcal{P}$ désigne l'ensemble de tous les motifs séquentiels. Un motif séquentiel $X = \langle X_1, \ldots, X_n \rangle$ est inclus dans un motif séquentiel $Y = \langle Y_1, \ldots, Y_m \rangle$, noté $X \sqsubseteq Y$, s'il existe n indices i_k pour $k \in 1, \ldots, n$ tels que $X_k = Y_{i_k}$ et $i_k < i_{k+1}$ pour $k \in 1, \ldots, n-1$. La concaténation de X et Y est définie par $X \cdot Y = \langle X_1, \ldots, X_n, Y_1, \ldots, Y_m \rangle$. Le support du motif est la proportion de séquences qui contiennent le motif : $supp(X) = \frac{|Seq \in \mathcal{S} | X \sqsubseteq Seq|}{|\mathcal{S}|}$. Un motif X est fréquent si son support dépasse un seuil défini par l'utilisateur. Soit α ce seuil, nous définissons $\mathcal{F}$ comme l'ensemble des motifs fréquents : $\mathcal{F} = \{X \in \mathcal{P} | supp(X) \geq \alpha\}$. Il est préférable de se limiter aux motifs fermés qui constituent une représentation sans perte d'information. Les motifs fréquents ne sont

ID	Séquences	Durées (en jours)
1	⟨**S**,**M**,**P**,A⟩	500
2	⟨**S**,B,**C**⟩	1,500
3	⟨**S**,D,E,**M**,**P**⟩	500
4	⟨**S**,**C**⟩	1,500
5	⟨**S**,**M**,F,G,**P**⟩	1,000
6	⟨**S**,**M**,**P**,**J**⟩	2,000
7	⟨**S**,**P**,**M**,**P**,**J**⟩	2,000
8	⟨**S**,H,**C**⟩	2,000

TAB. 1 – *Jeu de données jouet.*

Items	Significations
S	Hiring
C	Compensation
M	Mission
P	Pricing profile
J	Job
A,B,D,E,F,G,H	Other changes

TAB. 2 – *Changements.*

pas nombreux, mais nous utilisons toujours les motifs fermés car il n'y a aucun avantage réel à conserver deux motifs ayant exactement le même support avec l'un inclus dans l'autre. En considérant le seuil de support minimal $\alpha = 2/8$, les changements dans $\{S, C, M, P, J\}$ sont les seuls à contribuer à la présence de motifs fréquents, et ces changements conduisent à 4 motifs fermés fréquents : $\langle S \rangle$ (avec 8/8 de support), $\langle S, M, P \rangle$ (5/8), $\langle S, C \rangle$ (3/8) et $\langle S, M, P, J \rangle$ (2/8). Notez que les autres motifs fréquents ne sont pas fermés : par exemple, $\langle S, P \rangle$ n'est pas fermé puisqu'il existe le motif $\langle S, M, P \rangle$ de même support que $\langle S, P \rangle$ et incluant ce dernier.

Description de la courbe de survie de Kaplan-Meier La courbe de survie, couramment utilisée pour décrire la survenue de décès au fil du temps, représente la probabilité de survie en fonction du temps. Elle est généralement construite à l'aide de l'estimateur non paramétrique de Kaplan-Meier (E.L.Kaplan et Meier, 1958), basé sur les données de durée de vie. Cet estimateur divise la durée de participation observée en intervalles de temps, estimant la survie pour chaque intervalle tout en tenant compte des données censurées. La courbe de survie présente un aspect de *marche d'escalier* avec m marches, où m représente le nombre de moments où des décès sont observés, à ne pas confondre avec le nombre de décès (plusieurs décès peuvent survenir simultanément).

L'estimateur de Kaplan-Meier au temps t pour les individus ayant le motif X dans leur séquence, dénoté par $\widehat{Sur_X}(t)$, est défini comme suit :

$$\widehat{Sur_X}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i^X}{n_i^X}\right)$$

où t_i est le i^{th} temps où au moins un individu décède ; n_i^X est le nombre d'individus avec le motif X encore vivant juste avant le temps t_i ; et d_i^X est le nombre d'individus avec le motif X qui décèdent au temps t_i.

La figure 1 représente la visualisation de cette courbe de survie pour l'ensemble de la population (i.e., ayant le motif $\langle \emptyset \rangle$ dont la fermeture est $\langle S \rangle$) de notre exemple jouet. Tous les individus sont alignés à gauche car on utilise des durées et non pas des dates. Par conséquent, les employés toujours en poste seront considérés comme censurés à la fin de leur période d'observation qui se termine au moment du recueil des données. A chaque employé censuré, la valeur n_i est décrémentée ce qui conduit à une plus grande baisse de la courbe à la prochaine démission observée.

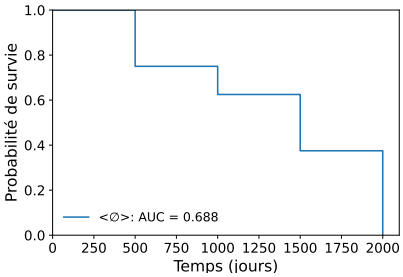

FIG. 1 – *Courbe de survie pour l'ensemble de la population du jeu de données jouet*

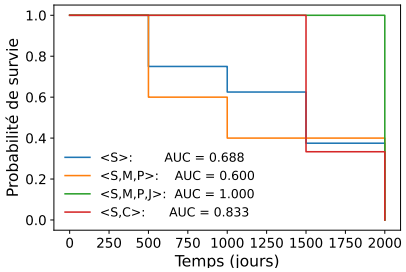

FIG. 2 – *Courbes de survie des motifs fréquents fermés du jeu de données jouet*

3 Motifs impactant la survie : définition et méthode

Vue d'ensemble de la méthode Dans cette étude, nous utilisons une méthodologie complète pour analyser les séquences et les facteurs qui contribuent à l'amélioration de la survie d'un motif séquentiel à l'autre. Notre approche étend l'exploration traditionnelle des motifs fréquents en tirant parti de l'estimateur de Kaplan-Meier. Plus précisément, nous commençons par identifier un motif contextuel, qui sert de point de départ à l'analyse de la survie. À partir de ce motif de contexte, nous cherchons à trouver les suffixes les plus simples qui modifient le taux de survie au-delà d'un seuil prédéfini.

Dans ce cas, les suffixes identifiés sont appelés *motifs impactant la survie* (voir ci-dessous). Le processus d'extraction comprend deux étapes : l'extraction de motifs séquentiels fréquents fermés et l'énumération des paires formées par un motif impactant la survie et son contexte. Par manque de place, nous ne détaillons pas l'algorithme dans cette version courte (voir Oubelmouh et al. (2023a)).

Cette approche nous permet d'identifier les motifs les plus influents et les plus exploitables qui ont un effet direct sur la probabilité de survie dans un contexte donné. En mettant l'accent sur la simplicité et la fréquence, nous pouvons identifier des motifs significatifs impactant la survie qui fournissent des informations précieuses pour prendre des décisions éclairées et mettre en œuvre des interventions ciblées.

Mesures d'intérêt pour l'analyse de survie Nous présentons les deux mesures d'intérêt originales de notre proposition. La première évalue l'intérêt d'un motif en mesurant son aire sous la courbe de survie (voir la définition 1). La seconde mesure prend en compte la différence entre la courbe de survie du motif évalué et celle d'un motif contextuel (voir la définition 2). La définition suivante formalise la notion de *AUC de survie* :

Definition 1 (AUC de survie) *L'aire normalisée sous la courbe de la fonction de survie de Kaplan-Meier pour un motif X est définie comme suit :*

$$AUC_{Sur}(X) = \frac{\int \widehat{Sur}_X(t)\,dt}{\Delta} = \frac{\sum_{i=0}^{m-1} \widehat{Sur}_X(t_i) \cdot (t_{i+1} - t_i)}{\Delta}$$

où Δ est la durée de l'employé qui est resté le plus longtemps dans l'ensemble de la population.

Intuitivement, la AUC de survie calcule l'aire sous la courbe de Kaplan-Meier en la normalisant entre 0 et 1 à l'aide de la durée maximale Δ. Globalement, plus l'AUC de survie d'un motif est élevée, plus la population couverte survit longtemps. Par exemple, l'AUC de survie pour le motif $\langle S \rangle$ correspond à l'AUC de survie de la courbe de Kaplan-Meier pour l'ensemble de la population : $AUC_{Sur}(\langle S \rangle) = \frac{1375}{2000} = 0,688$. Celle de $\langle S, M, P \rangle$ est légèrement moins bonne avec 0,60, tandis que celle de $\langle S, C \rangle$ est légèrement meilleure avec 0,83. En ce qui concerne notre jeu de données, la figure 2 montre les courbes de Kaplan-Meier des quatre motifs fréquents fermés avec leur AUC de survie.

L'AUC de survie est une mesure d'intérêt pertinente pour déterminer si un motif augmente ou diminue la survie. Dans le cas des changements que l'on sait négatifs, il est important de pouvoir y remédier en prenant des actions. Dans notre jeu de données jouet, les changements M et P aggravent la situation par rapport à S avec une réduction de $0.60 - 0.687 = -0.087$ de son *AUC de survie*. Il serait intéressant de savoir quels changements ultérieurs pourraient remédier à ce déclin. Afin d'identifier de tels motifs, nous proposons de mesurer la différence entre la *AUC de survie* du motif X et celle correspondant au motif de référence, appelé *motif contextuel* (ou contexte en abrégé) :

Definition 2 (Gain de survie) *Compte tenu d'un motif contextuel C, le gain de survie du motif X correspond à la différence entre l'AUC de survie de $C \cdot X$ et celle de C : $Gain_C(X) = AUC_{Sur}(C \cdot X) - AUC_{Sur}(C)$.*

Il est facile de voir que le gain de survie est une mesure dont la valeur est comprise entre -1 et 1. Un gain de survie positif (resp. négatif) signifie que X améliore (resp. détériore) le taux de survie. Par exemple, comme indiqué précédemment, nous constatons que $Gain_{\langle S \rangle}(\langle M, P \rangle) = -0,087$ soulignant l'impact négatif de $\langle M, P \rangle$ dans le contexte $\langle S \rangle$. Inversement, le motif $\langle C \rangle$ améliore le taux de survie comme $Gain_{\langle S \rangle}(\langle C \rangle) = AUC_{Sur}(\langle S, C \rangle) - AUC_{Sur}(\langle S \rangle) = 0.833 - 0.687 = 0.146$.

Plus le gain de survie d'un motif est éloigné de zéro, plus ce modèle est pertinent. C'est pourquoi la définition suivante formalise la notion de *motifs impactant la survie* :

Definition 3 (Motifs impactant la survie) *Un motif X est un motif impactant la survie pour le contexte C si son gain de survie en valeur absolue est supérieur à un seuil σ spécifié par l'utilisateur : $|Gain_C(X)| > \sigma$.*

Illustrons cette définition par notre exemple. Les motifs $\langle M, P \rangle$ et $\langle C \rangle$ agissent respectivement comme des *motifs impactant la survie négatif* et *positif* dans le même contexte $\langle S \rangle$ pour le seuil $\sigma = 0,05$. Il est intéressant de noter qu'il est possible qu'un *motif impactant la survie* négatif soit à son tour le contexte d'un *motif impactant la survie* positif. Par exemple, comme nous avons $Gain_{\langle S,M,P \rangle}(\langle J \rangle) = 1 - 0,687 = 0,313$, le modèle $\langle J \rangle$ est un *motif impactant la survie* positif pour le contexte $\langle S, M, P \rangle$. Dans le cas de l'attrition des employés, cela signifie que si les événements $\langle M, P \rangle$ ne peuvent pas être évités, il est alors pertinent de considérer $\langle J \rangle$ afin de conserver l'employé. De tels motifs fournissent évidemment des informations très précieuses.

4 Expérimentations sur nos données

Dans notre article original (Oubelmouh et al., 2023a), la section expérimentale évalue la quantité et le type de motifs séquentiels pertinents pour la survie et répond aux questions clés suivantes :

— Quelle est l'influence du seuil de fréquence α sur les résultats des algorithmes ?
— Quelle est l'influence du seuil de gain σ sur nos résultats ?
— Comment combiner les deux seuils pour obtenir les meilleurs résultats ?

L'évaluation de l'efficacité de notre approche n'entre pas dans le cadre de cet article où nous nous concentrons plutôt sur un ensemble de données réelles (où toutes les expériences ont été réalisées en quelques minutes). Dans le cadre de ce résumé, nous allons uniquement répondre à la dernière question.

Description des données La description de nos données est présente dans l'article original publié à DSAA (Oubelmouh et al., 2023a) ainsi que dans l'article publié à l'atelier GAST d'EGC (Oubelmouh et al., 2023b). Au total nous avons utilisé 7,527 séquences qui en moyenne contiennent 5,25 événements. La plus petite est de longueur 2 et la plus grande de longueur 26. Nous allons juste décrire les 7 changements considérés dans nos séquences, qui sont les suivantes : S indique l'embauche, C représente un changement de rémunération, M indique un changement de mission, P indique un changement de pricing profile (profil vendu aux clients, e.g. « Consultant Junior », « Consultant Senior »), $J2$ représente un changement de poste, $J3$ indique un changement de titre et $J4$ indique un transfert au sein de l'entreprise.

Interprétation des résultats dans le contexte de l'attrition Les figures 3 et 4 présentent deux exemples de successions de *motifs impactant la survie* en fonction de deux gains de survie différents (0.01 et 0.05). Nous pouvons voir qu'avec un gain de survie plus élevé, il est nécessaire d'avoir un *motif impactant la survie* plus efficace et donc souvent le motif est plus complexe.

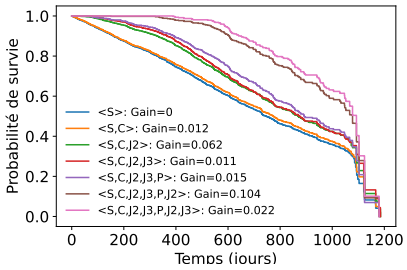

FIG. 3 – *Courbes de survie pour des motifs successifs avec* $\sigma = 0.01$

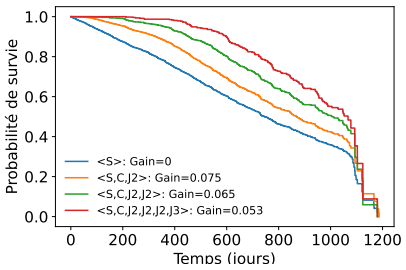

FIG. 4 – *Courbes de survie pour des motifs successifs avec* $\sigma = 0.05$

Les résultats de notre étude sont particulièrement pertinents dans le contexte de l'attrition des employés au sein d'une organisation. Lorsque l'on cherche à améliorer les perspectives de survie d'un employé sur la base de son motif contextuel, il devient

crucial d'identifier un *motif impactant la survie* qui soit non seulement très efficace en termes de gain, mais aussi gérable en termes de complexité. La mise en œuvre d'une multitude d'événements dans un court laps de temps pour un seul employé peut s'avérer difficile, voire impossible, comme l'application d'un même événement à plusieurs reprises.

Nous avons ensuite tracé les boîtes à moustaches représentant les 10 motifs impactant la survie qui ont été extraits le plus de fois par notre algorithme avec un seuil de gain de 0.05, en haut de chaque boîte à moustaches est inscrit ce nombre.

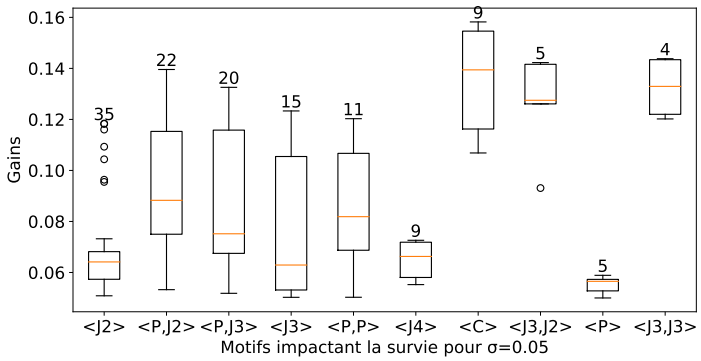

FIG. 5 – *Boîtes à moustaches du top 10 des motifs impactant la survie quand* $\sigma = 0.05$

Sans surprise, on observe que le *motifs impactant la survie* avec le meilleur gain en moyenne est le pattern $\langle C \rangle$ représentant un changement de salaire. La figure 5 montre qu'en moyenne, les *motifs impactant la survie* dans le top 10 sont légèrement plus complexes (comparé aux boîtes à moustaches pour le seuil à 0.01 non présenté ici), ce qui suggère que l'obtention d'un gain significatif dans l'*AUC de survie* pour un contexte donné peut nécessiter plusieurs événements plutôt qu'un seul. Prenons par exemple le *motif impactant la survie* $\langle P, J2 \rangle$ (changement de pricing profile et changement de poste), qui est le deuxième *motif impactant la survie* le plus fréquent de la collection. Ce motif permet à lui seul à au moins 22 motifs de contexte différents d'améliorer l'AUC de survie de plus de 0.05.

5 Conclusion

Dans cette étude, nous avons présenté une nouvelle approche qui contribue à la compréhension des dynamiques d'attrition au sein des organisations en fournissant une méthodologie permettant d'identifier les *motifs impactant la survie* basés sur le contexte des employés. Ces motifs permettent de comprendre les séquences d'événements qui influencent de manière significative la probabilité de survie d'un employé. En donnant la priorité aux motifs présentant un gain élevé et une complexité gérable, les organisations peuvent développer des interventions ciblées pour améliorer la rétention des employés et réduire les taux d'attrition. Toutefois, l'application de ces résultats

doit être adaptée au contexte, en tenant compte des caractéristiques et des contraintes propres à l'organisation et à son personnel.

Remerciements. Ce travail a été partiellement financé par le programme CIFRE (ANRT 2021/0760).

Références

E.L.Kaplan et P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association 53*(282), 457–481.

Mattos, J. B. (2021). A supervised descriptive local pattern mining approach to the discovery of subgroups with exceptional survival behaviour. Master's thesis, Universidade Federal de Pernambuco.

Mooney, C. H. et J. F. Roddick (2013). Sequential pattern mining–approaches and algorithms. *ACM Computing Surveys (CSUR) 45*(2), 1–39.

Oubelmouh, Y., F. Fargon, C. de Runz, A. Soulet, et C. Veillon (2023a). Identifying survival-changing sequential patterns for employee attrition analysis. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*.

Oubelmouh, Y., F. Fargon, C. de Runz, A. Soulet, et C. Veillon (2023b). Le temps, un challenge à prendre en considération dans l'attrition des employés. In *Atelier Gestion et Analyse de données Spatiales et Temporelles@ EGC2023*, pp. 1–13.

Ren, Y., K. Zhang, et Y. Shi (2019). Survival prediction from longitudinal health insurance data using graph pattern mining. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1104–1108. IEEE.

Ritschard, G., A. Gabadinho, N. S. Muller, et M. Studer (2008). Mining event histories : A social science perspective. *International Journal of Data Mining, Modelling and Management 1*(1), 68–90.

Truong-Chi, T. et P. Fournier-Viger (2019). *A Survey of High Utility Sequential Pattern Mining*, pp. 97–129. Cham : Springer International Publishing.

Wang, P., Y. Li, et C. K. Reddy (2019). Machine learning for survival analysis : A survey. *ACM Computing Surveys (CSUR) 51*(6), 1–36.

Summary

This work is a short version of (Oubelmouh et al., 2023a) that addresses the issue of employee attrition within organizations. Whereas reducing attrition has become a major objective for compagnies, the literature on this topic is relatively limited compared with that on customer attrition. Moreover, even studies addressing this topic often fail to take into account the impact of time and duration on attrition rates. This research combines two approaches to fill this gap: the exploration of frequent patterns in event sequences and survival analysis with Kaplan-Meier. This study identifies events with a significant impact on the survival estimator, named survival-changing sequential patterns. The results suggest that some patterns and the addition of specific events can positively or negatively influence employee retention.